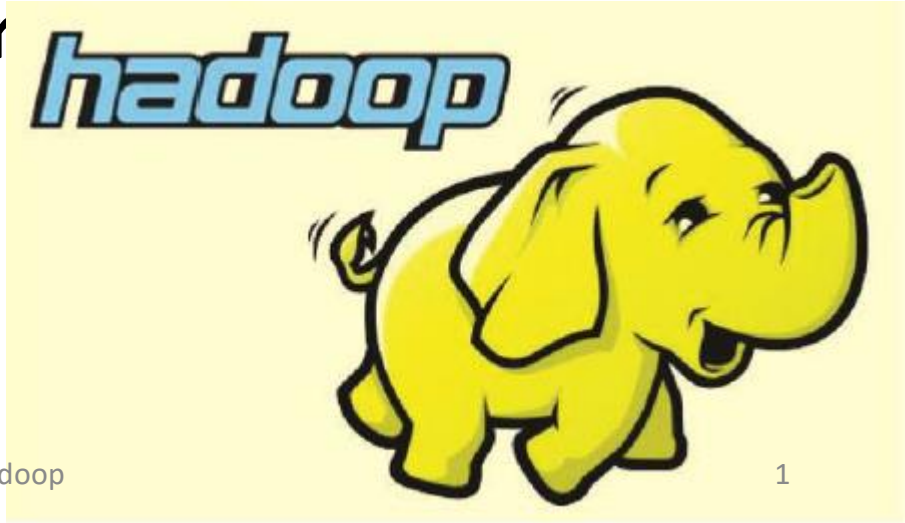
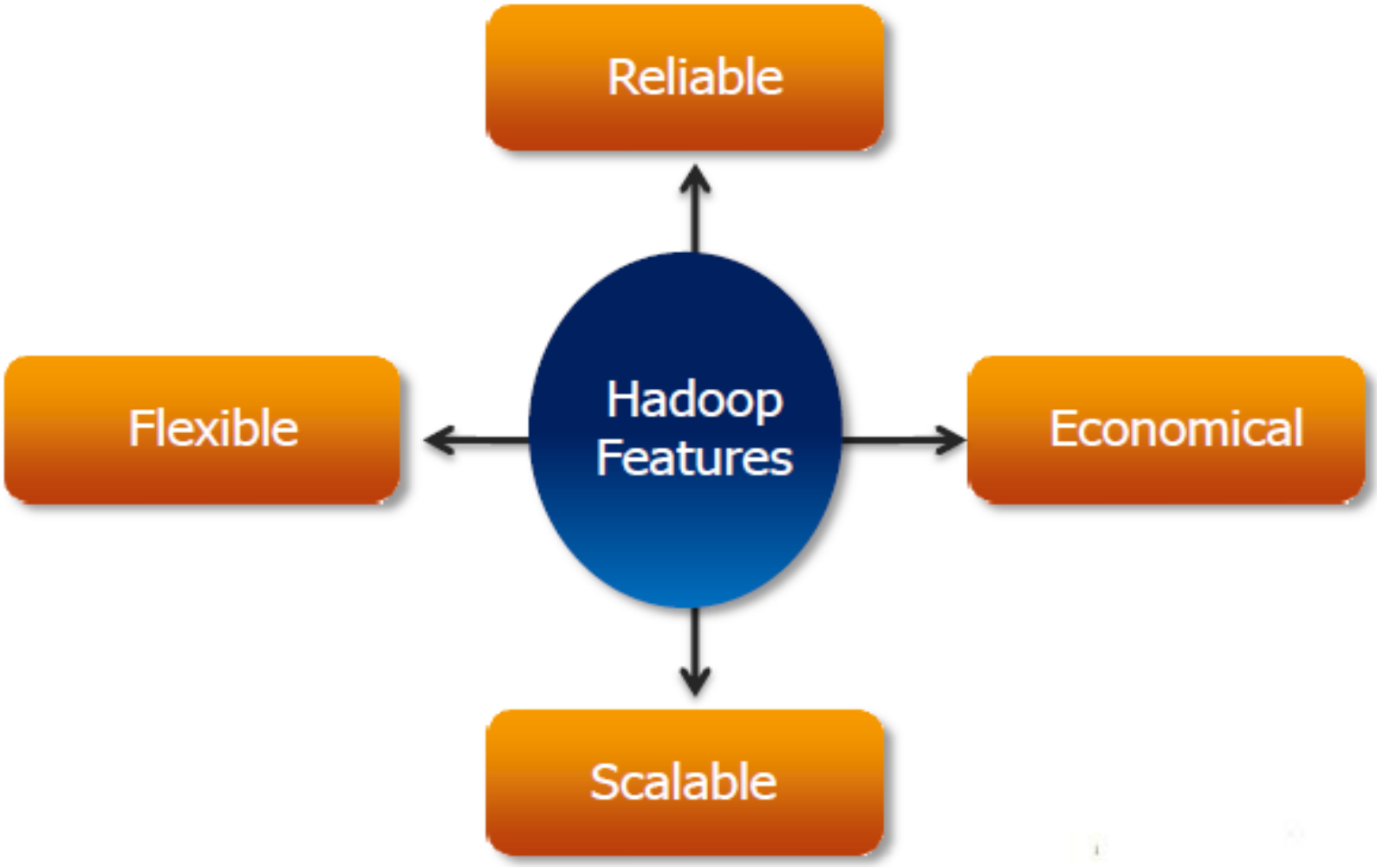
What is Hadoop ?

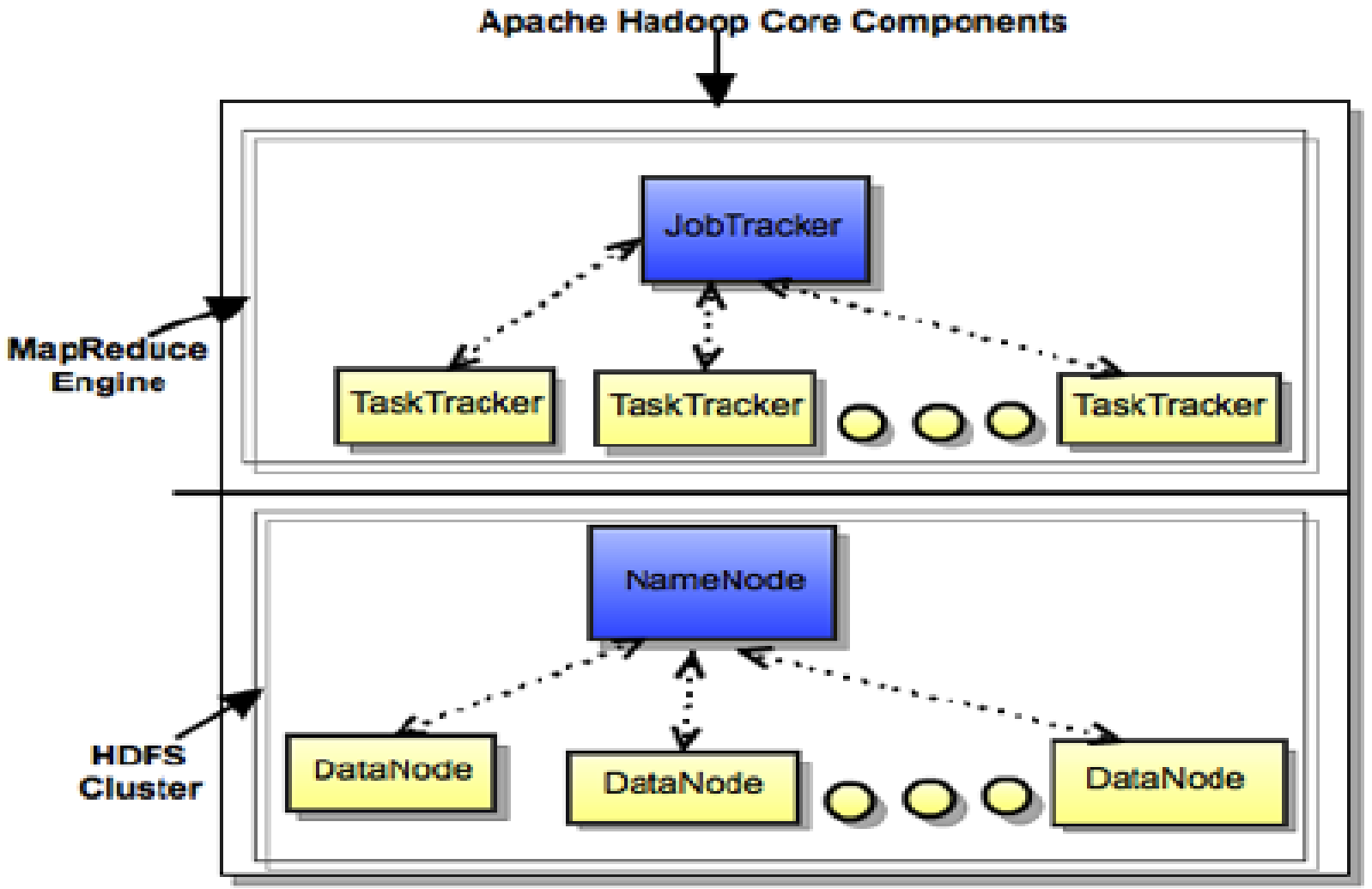
- Apache Hadoop is a **framework** **that allows for the distributed processing of large data sets across clusters of commodity computers using a simple programming model.**
- **Open-source Data Management with scale-out storage & distributed pr**



Hadoop – Key Characteristics



Hadoop Core Components



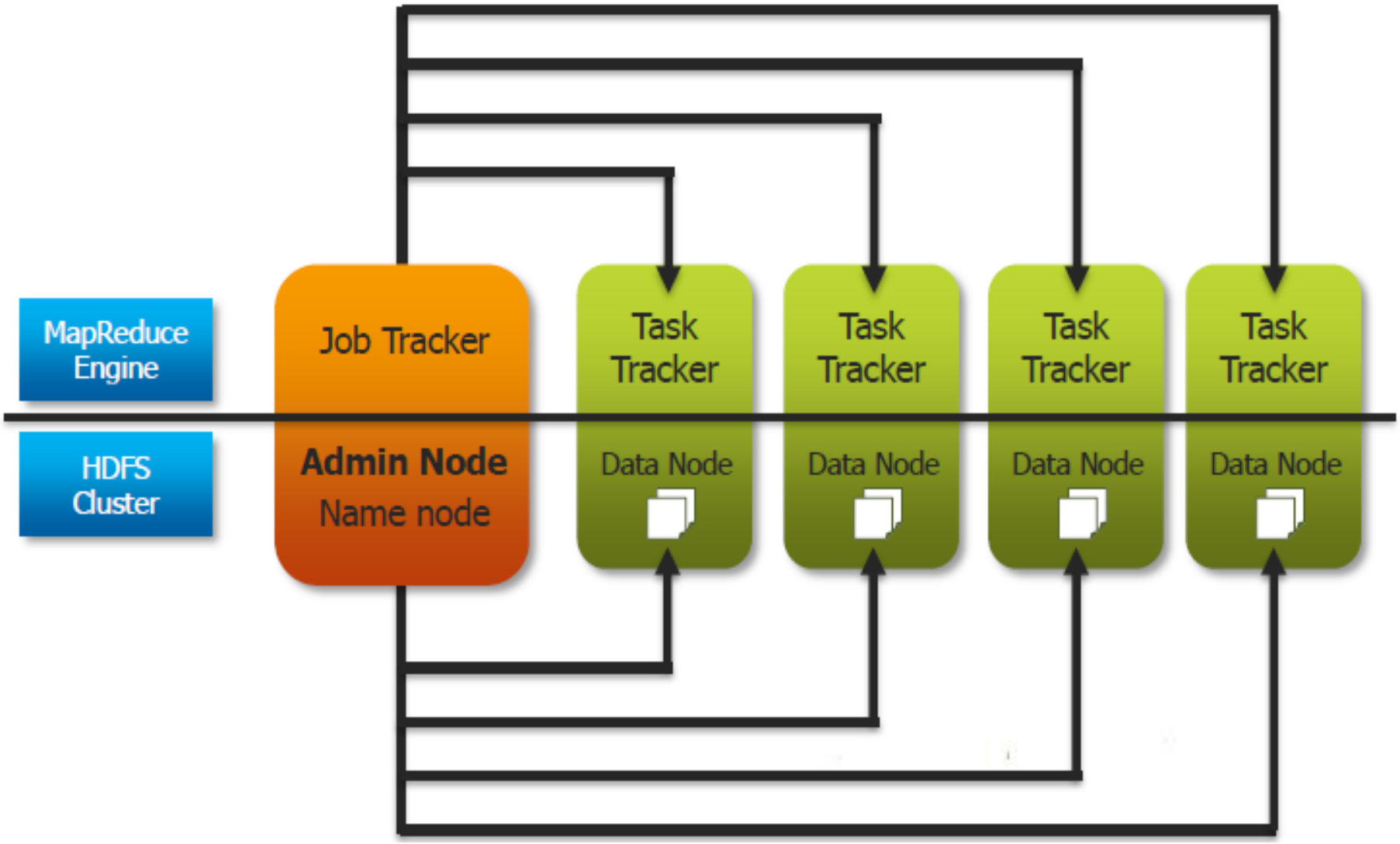
Hadoop Core Components

Hadoop is a system for large scale data processing.

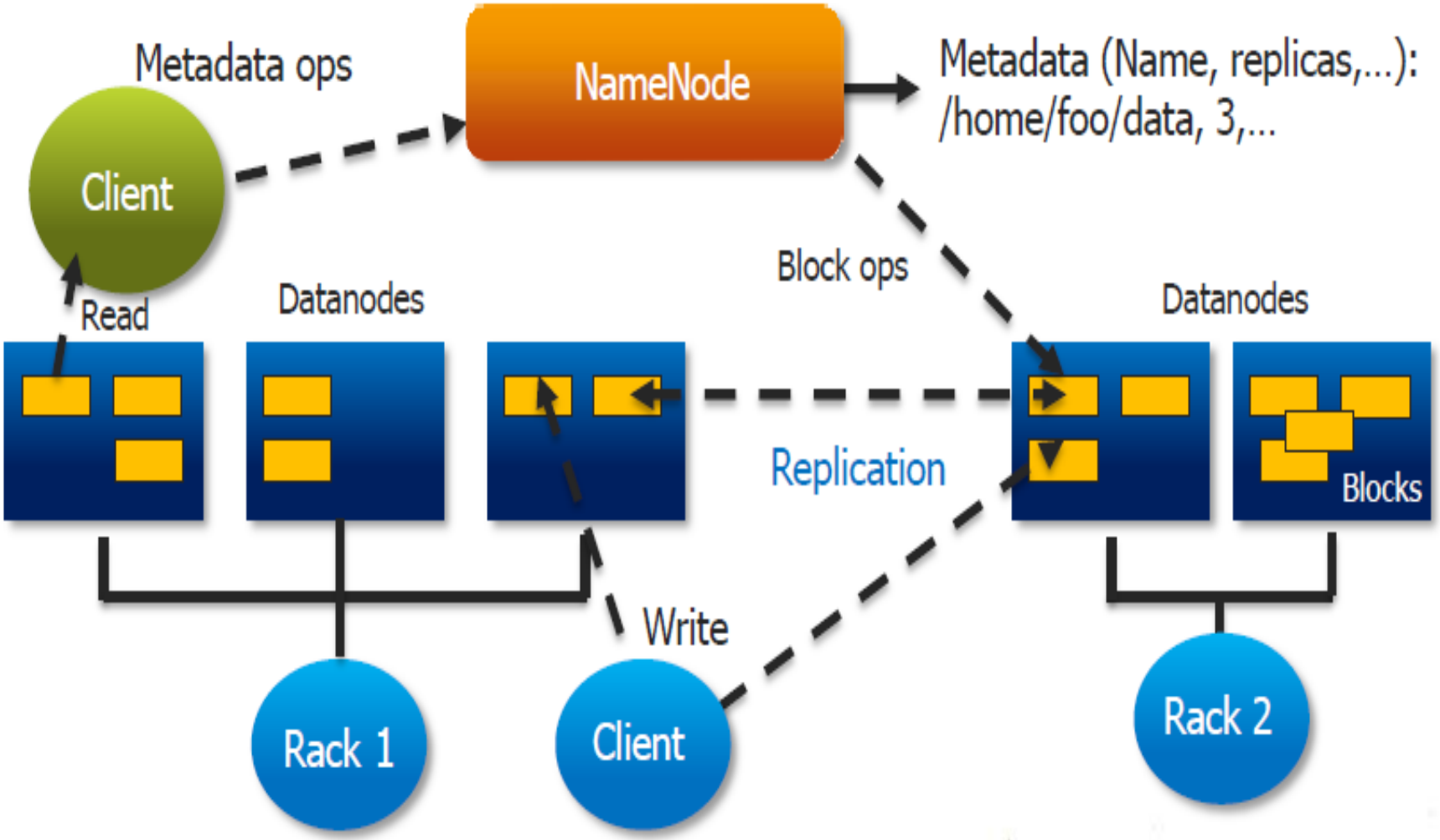
It has two main components:

- ✓ **HDFS – Hadoop Distributed File System (Storage)**
  - ✓ Distributed across "nodes"
  - ✓ Natively redundant
  - ✓ NameNode tracks locations.
- ✓ **MapReduce (Processing)**
  - ✓ Splits a task across processors
  - ✓ "near" the data & assembles results
  - ✓ Self-Healing, High Bandwidth
  - ✓ Clustered storage
  - ✓ JobTracker manages the TaskTrackers

Hadoop Core Components (cont..)



HDFS Architecture



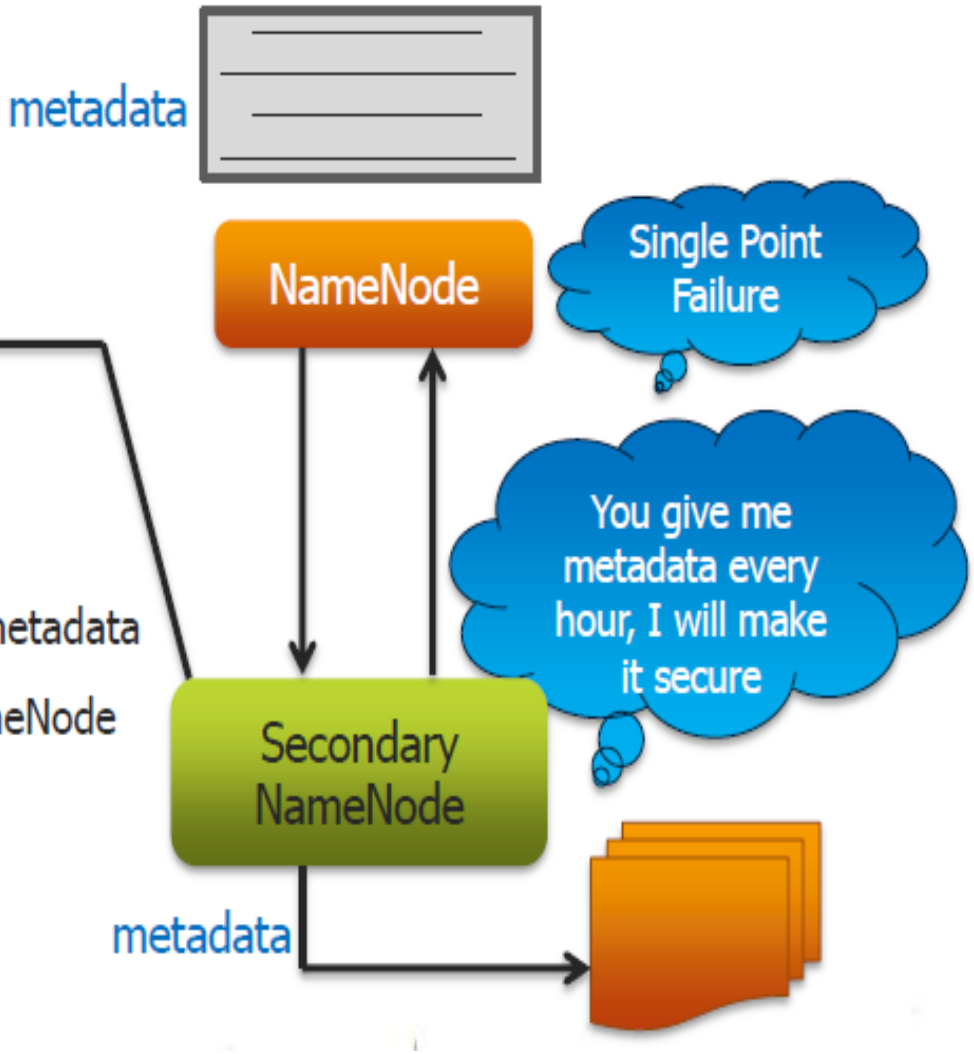
Main Components of HDFS

- ✓ NameNode:
  - ✓ master of the system
  - ✓ maintains and manages the blocks which are present on the DataNodes

- ✓ DataNodes:
  - ✓ slaves which are deployed on each machine and provide the actual storage
  - ✓ responsible for serving read and write requests for the clients

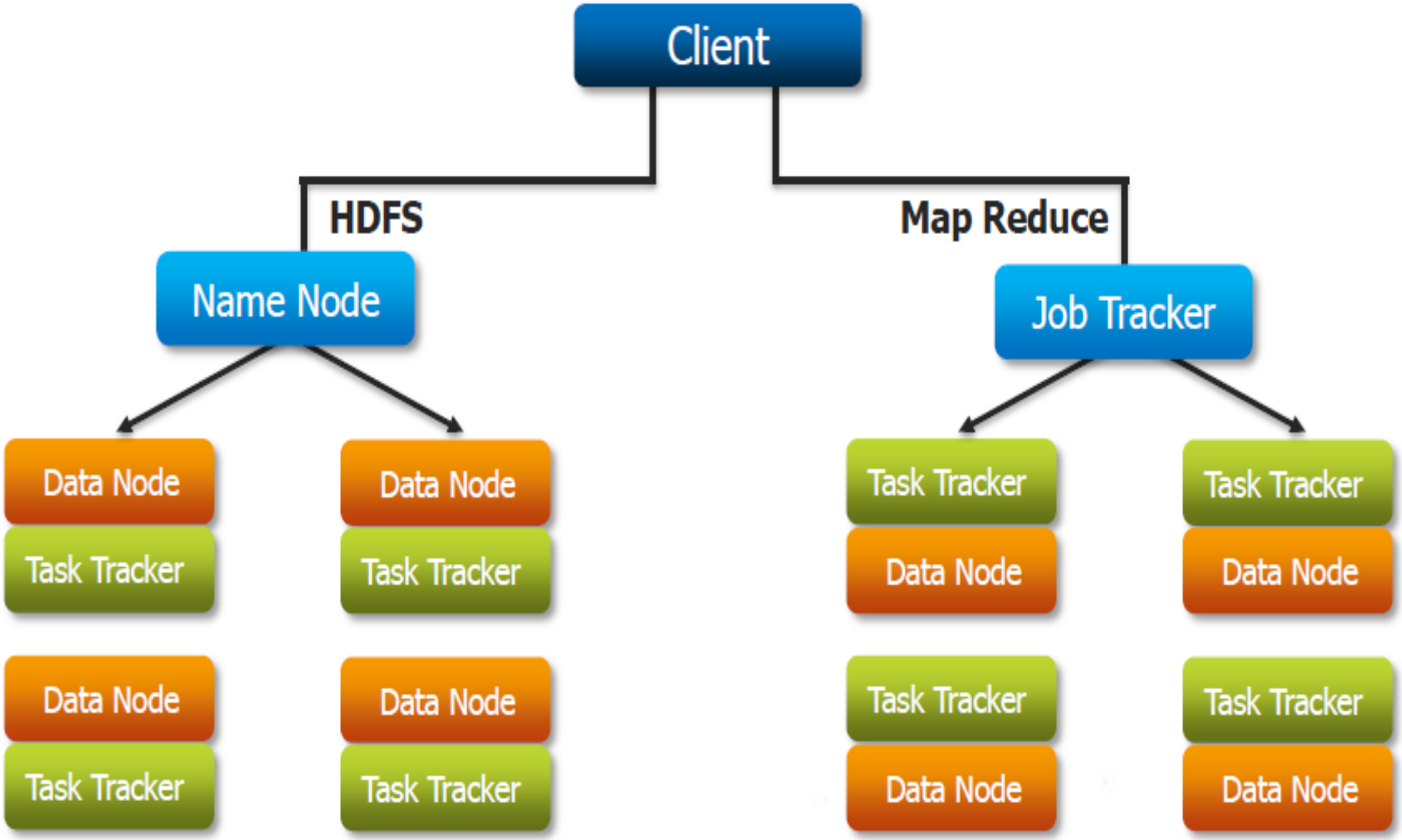
Secondary Namenode

- ✓ Secondary NameNode:
  - ✓ Not a hot standby for the NameNode
  - ✓ Connects to NameNode every hour*
  - ✓ Housekeeping, backup of NemeNode metadata
  - ✓ Saved metadata can build a failed NameNode

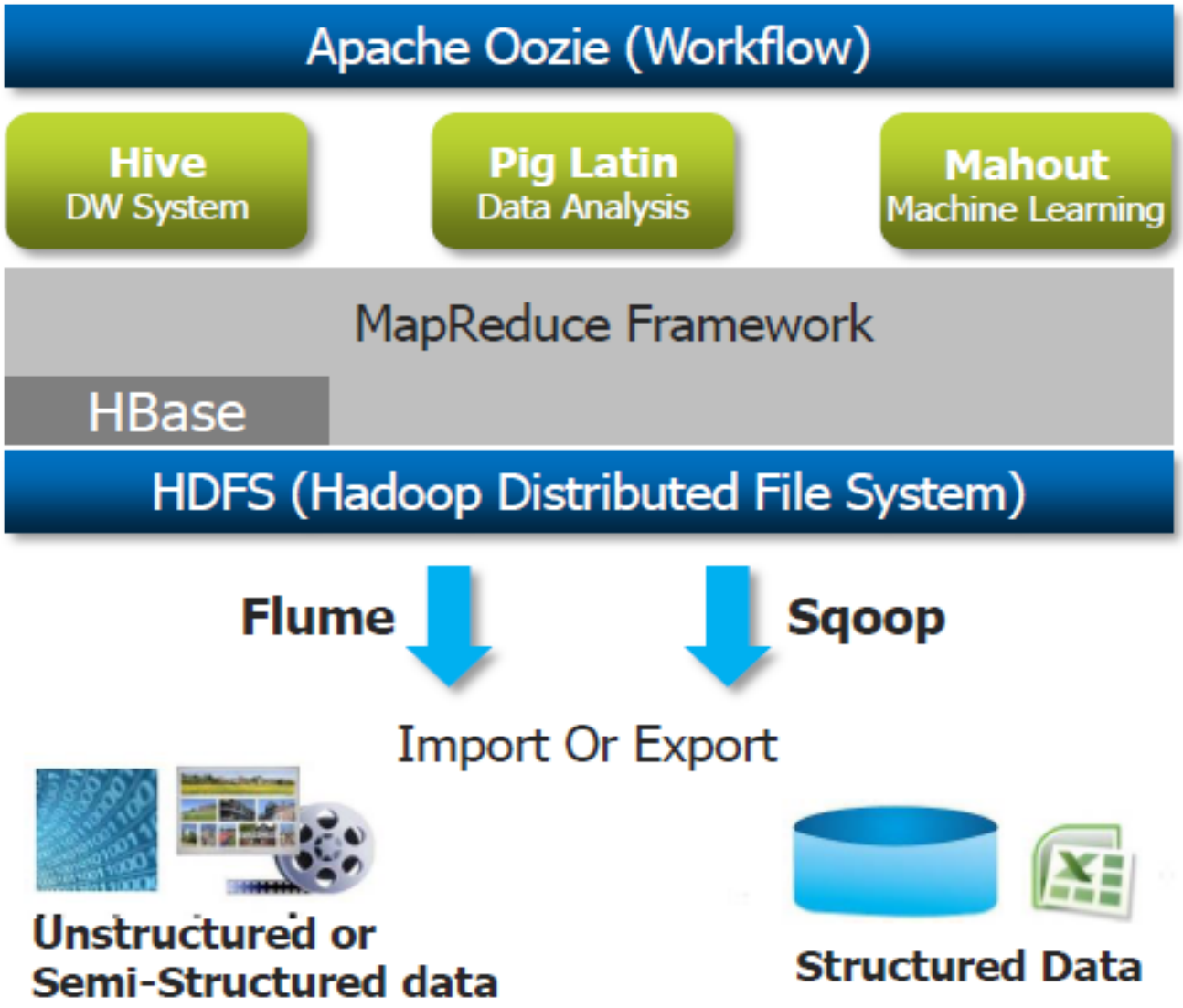


Hadoop Cluster Architecture

Hadoop Cluster Architecture



Hadoop Ecosystem



Hadoop Cluster Modes

Hadoop can run in any of the following three modes:

Standalone (or Local) Mode

- ✓ No daemons, everything runs in a single JVM.
- ✓ Suitable for running MapReduce programs during development.
- ✓ Has no DFS.

Pseudo-Distributed Mode

- ✓ Hadoop daemons run on the local machine.

Fully-Distributed Mode

- ✓ Hadoop daemons run on a cluster of machines.

Nowadays…

- When you visit yahoo, you are interacting with data processed with Hadoop!

Existing Hadoop Customers

Retail

- CRM – Customer Scoring
- Store Siting and Layout
- Fraud Detection / Prevention
- Supply Chain Optimization

DataSong

Advertising & Public Relations

- Demand Signaling
- Ad Targeting
- Sentiment Analysis
- Customer Acquisition

Financial Services

- Algorithmic Trading
- Risk Analysis
- Fraud Detection
- Portfolio Analysis

Morgan Stanley

Media & Telecommunications

- Network Optimization
- Customer Scoring
- Churn Prevention
- Fraud Prevention

Manufacturing

- Product Research
- Engineering Analytics
- Process & Quality Analysis
- Distribution Optimization

AA AmericanAirlines

Energy

- Smart Grid
- Exploration

Government

- Market Governance
- Counter-Terrorism
- Econometrics
- Health Informatics

AADHAAR

Healthcare & Life Sciences

- Pharmaco-Genomics
- Bio-Informatics
- Pharmaceutical Research
- Clinical Outcomes Research

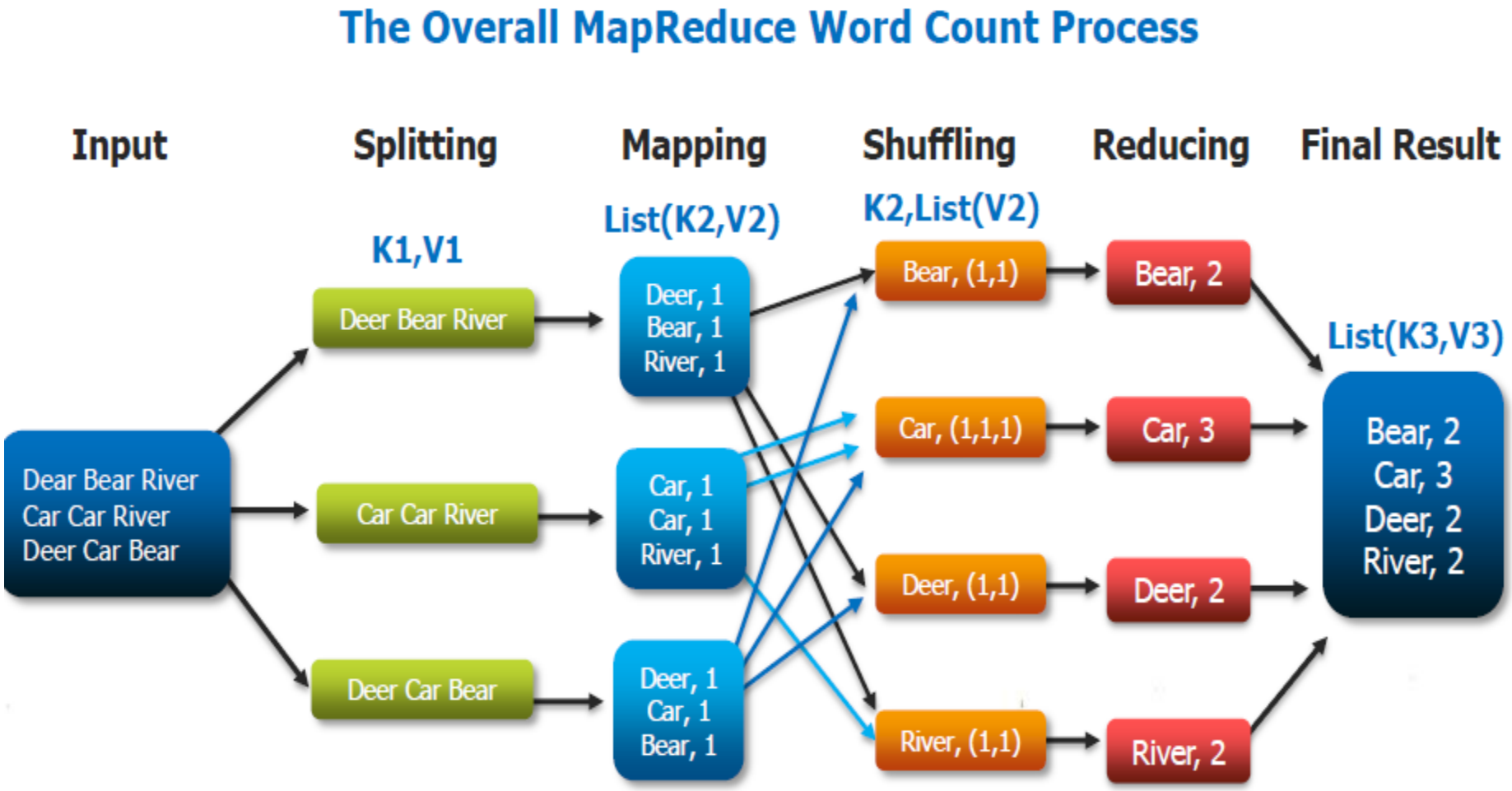
GE Healthcare

Hadoop vs Spark

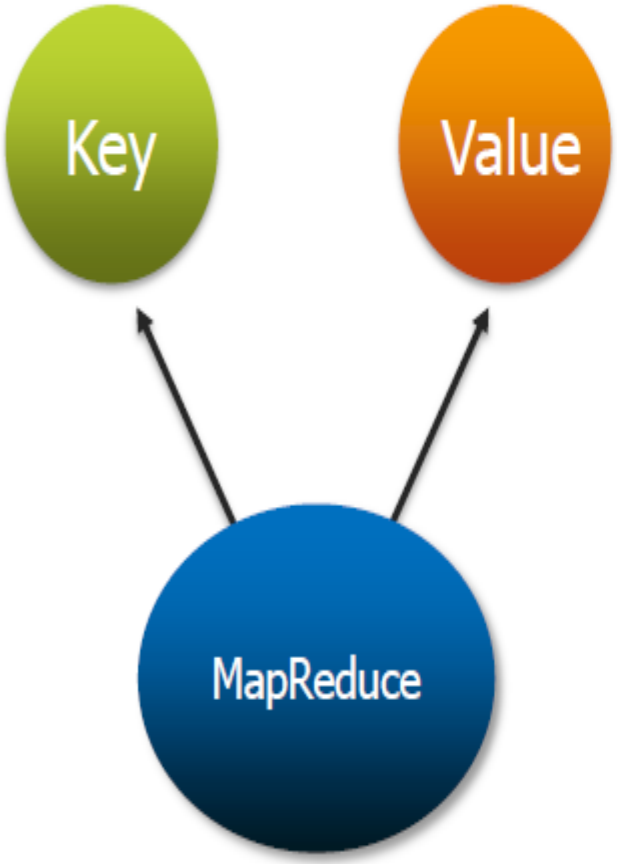
HADOOP	SPARK
Stores data on disk	Sores data in memory (RAM)
Commodity hardware can be utilized	Need high end systems with greater RAM
Uses Replication to achieve fault tolerance	Uses different data storage models to achieve fault tolerance (Eg. RDD)
Speed of processing is less due to disk read write	100x faster than Hadoop
Supports only Java & R	Supports Java, Python, R, Scala etc. Ease of programming is high.
Everything is just Map and Reduce	Supports Map, Reduce, SQL. Streaming etc
Data should be in HDFS	Data can be in HDFS,Cassandra,Hbase or S3. Runs on Hadoop, Cloud, Mesos or standalone

- https://www.mentimeter.com/

Map Reduce Programming



Anatomy of a MR Program



Map:

(K1, V1) → List (K2, V2)

Reduce:

(K2, list (V2)) → List (K3, V3)

Map Reduce Flow

Input data is distributed to nodes

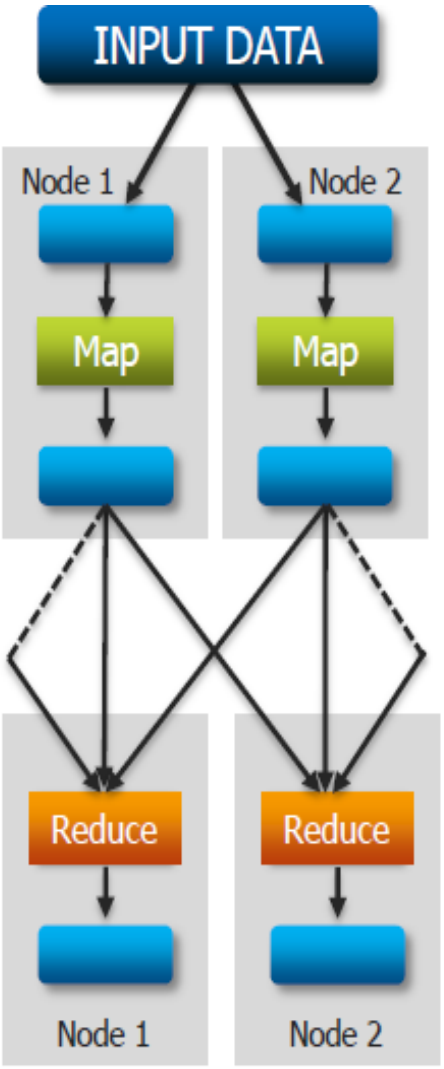
Each map task works on a “split” of data

Mapper outputs intermediate data

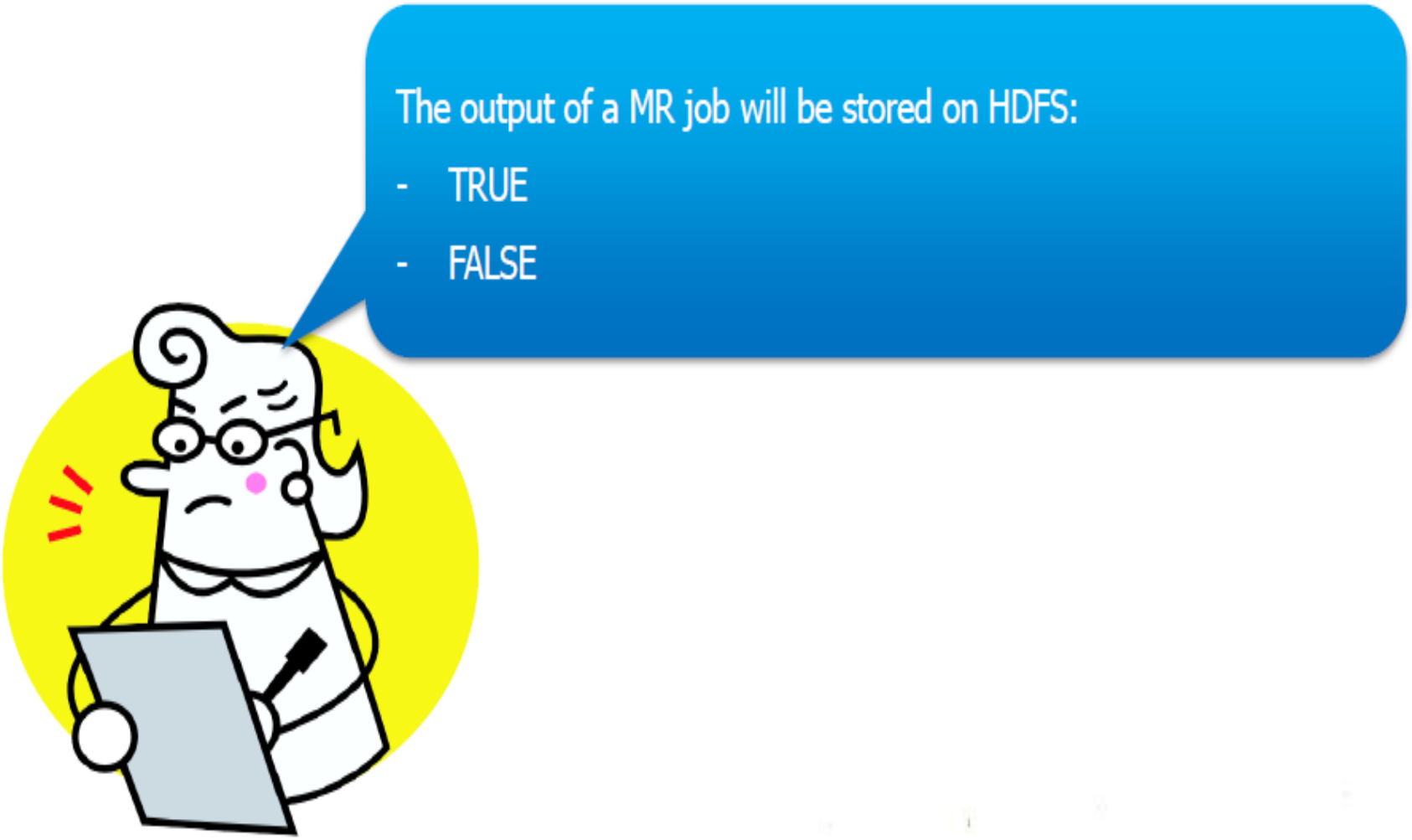
Data exchange between nodes in a “shuffle” process

Intermediate data of the same key goes to the same reducer

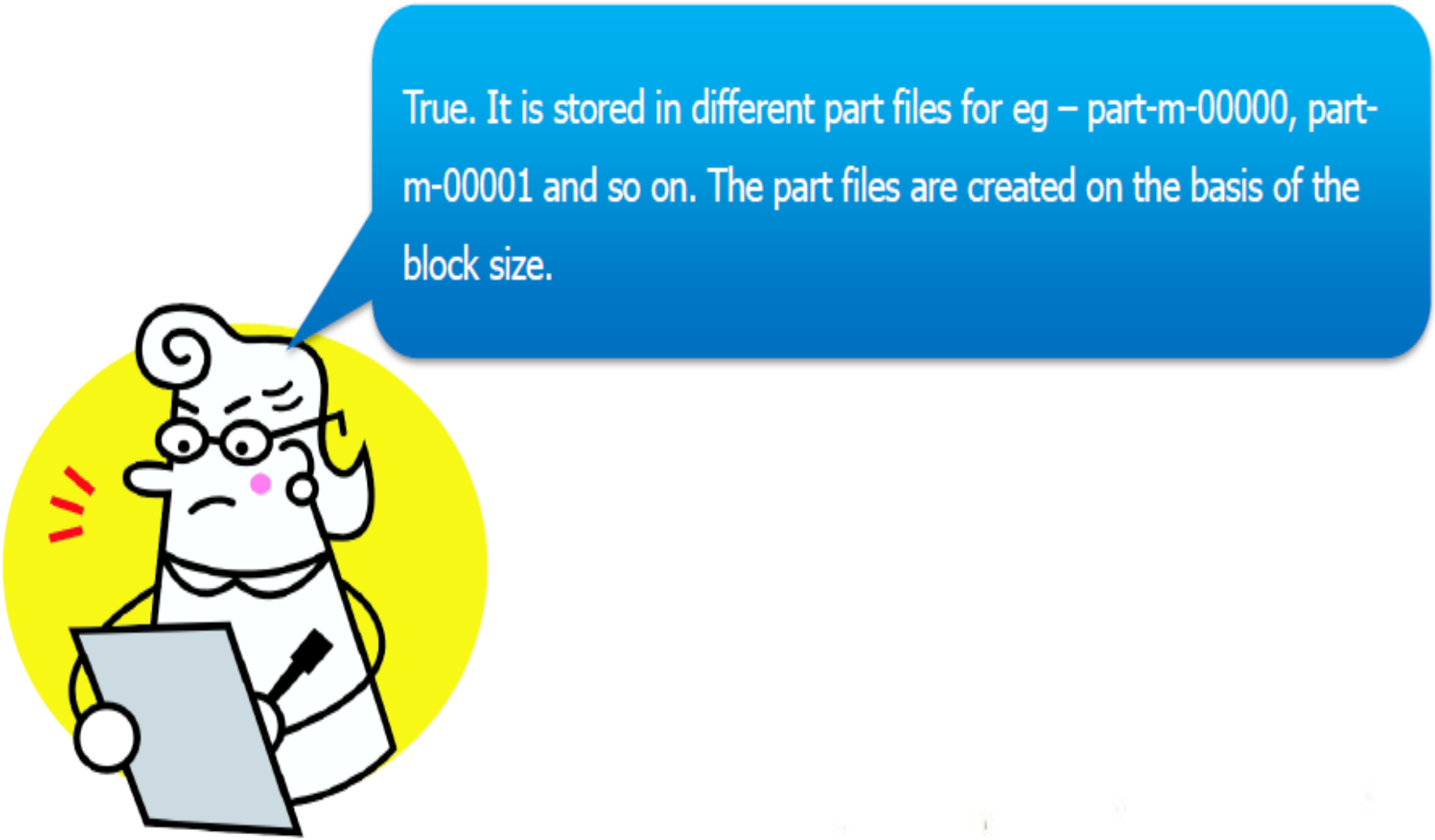
Reducer output is stored



Annie's Question



Annie's Answer

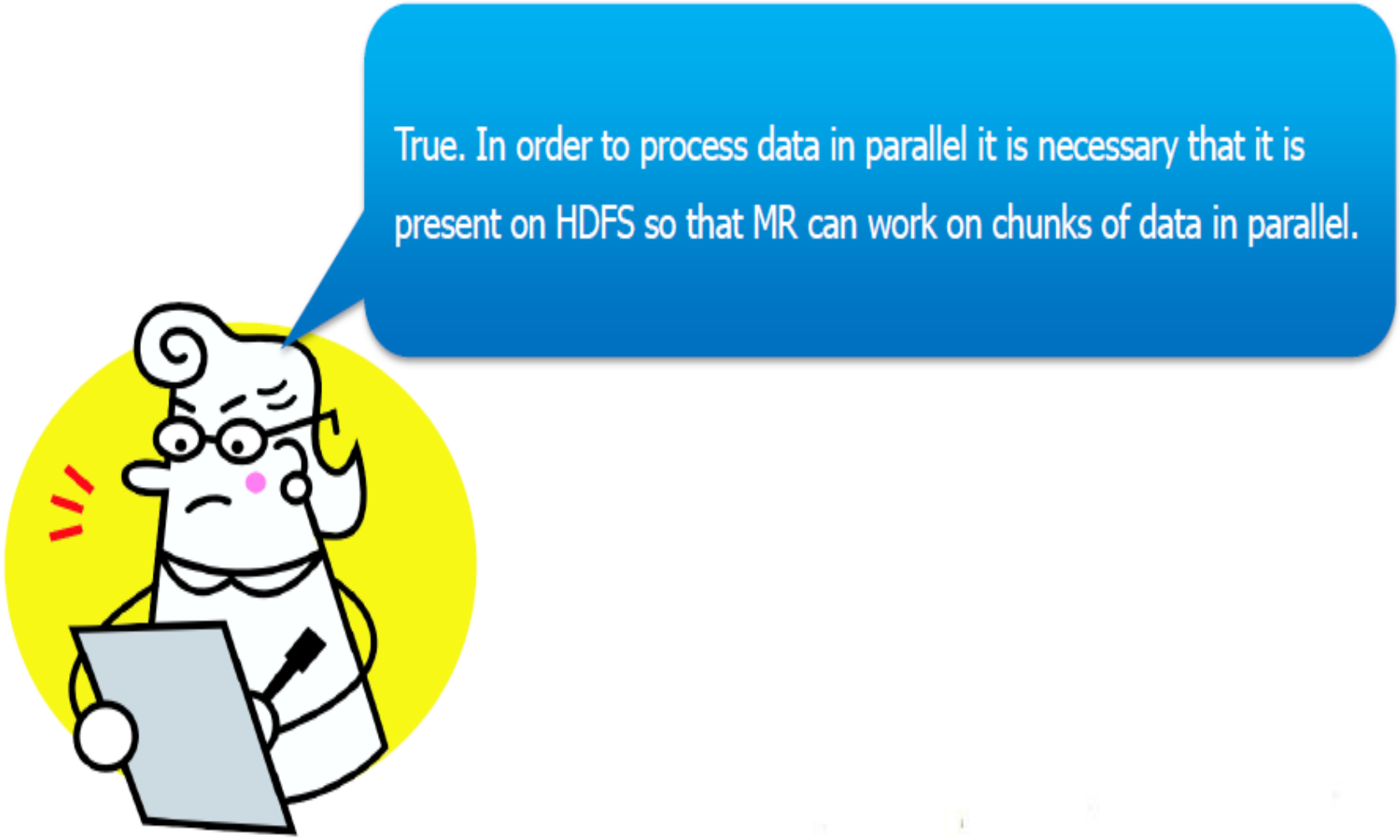


Annie's Question

To run MR job data should be present on HDFS:

- TRUE
- FALSE

Annie's Answer



Slides Credit

- Lija Mohan, CUSAT